Optimal Batch Variance with Second-Order Marginals

Zelda Mariet *Google* zmariet@google.com
Joshua Robinson *MIT LIDS & CSAIL* joshrob@mit.edu
Jamie Smith *Google* jamieas@google.com
Suvrit Sra *MIT LIDS* suvrit@mit.edu
Stefanie Jegelka *MIT CSAIL* stefje@csail.mit.edu

Abstract

Obtaining unbiased, low-variance estimates of the mean of a ground set of points by sampling a small subset of points is a crucial machine learning task, arising in stochastic learning procedures, experimental design, and active learning. In the purely stochastic case, it is well-known that importance sampling achieves unbiased estimates of minimal variance; but finding optimal distributions over batches of size greater than 1 has proven difficult. For batches of arbitrary size, we show a) that importance sampling achieves the lowest variance when sampling *with* replacement, and b) that parametric distributions over batches can be optimized via a quadratic form in the distribution's first and second order marginals when sampling *without* replacement. We verify that such learned distributions outperform other batch selection methods, and achieve faster SGD convergence in downstream experiments. As a side-effect of our analysis, we show that distributions over fixed-sized subsets cannot be characterized by their first- and second- order marginals in polynomial time.

1. Introduction and related work

Most modern machine learning models that are learned using large quantities of data rely on first-order gradient methods for training; typically, the gradient update rule is computed using an *estimate* of the gradient, as computing the full-gradient would be too computationally intensive. Most commonly, the estimated gradient is obtained using a batched stochastic estimate, where the selected batch of training points is sampled uniformly at random, without replacement, from the training set.[1]

Common improvements upon stochastic gradient methods leverage variance reduction to reduce the variance $\mathbb{E}\left[\|\nabla\bar{f}\|^2\right]$ of the gradient estimate $\nabla\bar{f}$ [9, 20, 22, 6]. In the purely stochastic setting (batch size $k = 1$), the estimated gradient variance is minimized by sampling training points from the importance sampling distribution [18, 1, 3]; however, extending this result to larger batch sizes is a non-trivial endeavor.

Subset-selection for unbiased and low-variance estimates is not limited to applications to stochastic gradient methods: this question appears as a fundamental optimization problem with crucial applications to a variety of machine learning problems, such as active learning [7, 8] and, more generally, experimental design [5, 19, 11].

Experimental design and active learning applications typically focus on finding an optimal *set* of points to evaluate; motivated by batched stochastic applications, we focus in this paper on designing *distributions* over fixed-size subsets. In this paper we study the question: *how can you sample a minibatch giving a minimal variance and unbiased estimate?*

1. This is commonly implemented by applying a random permutation of the training data, then sampling batches of points sequentially from the permutation.

For the purpose of our analysis and experiments, we make the strong assumption that we can query the norms $\|x_i\|$ and inner products $x_i^\top x_j$ of the ground set of points, as in [1, 3]. Standard tricks to amortize the cost of querying these values can be used downstream for batched stochastic SGD applications, such as using stale computations and Lipschitz upper bounds approximations [24, 15, 10].

Contributions. When sampling subsets to reconstruct unbiased mean estimates, we show:
- Distribution optimality can be characterized by first and second order marginals
- Constructing a distribution by first/second-order marginals is NP-hard in the general case
- When sampling points independently and sequentially, importance sampling (S) is optimal
- Under simplifying assumptions, IS generalizes to sampling without replacement
- Experimentally, using first and second-order marginals to learn a parametric distribution improves variance reduction and faster SGD convergence in downstream experiments.

Notation. We consider sampling both with and without replacement. For a ground set $\mathcal{Y}$ of n items, we write $\mathcal{P}_{n,k}$ the set of distributions over multisets S of $\mathcal{Y}$ of size k (where S may contain repeated items). Similarly, we write $\mathcal{Q}_{n,k} \subset \mathcal{P}_{n,k}$ the set of distributions over subsets S of $\mathcal{Y}$ with support on sets of size exactly k (where all elements of S are distinct).

2. Variance reduction and pairwise counts

We consider a ground set of n vectors $\mathcal{Y} = \{x_1, \dots, x_n\}$, where each x_i is an element of $\mathbb{R}^d$; we seek to approximate the mean vector $\mu = \frac{1}{n}\sum_i x_i$. Unless mentioned otherwise, batches are sampled *with* replacement, and so all sets should be understood to be multisets. The standard stochastic approach to estimating $\bar{x}$ simply samples a subset S of k points uniformly at random; then, μ can be approximated as

$$\mu_{\text{unif}}(S) = \frac{1}{k}\sum\nolimits_{i\in S} x_i. \tag{1}$$

More generally, for any distribution $p \in \mathcal{P}_{n,k}$, we can construct an unbiased estimate of μ by reweighting sampled points by their expected sampling frequency. Let $c_i(S)$ represent the number of times point i appears in a subset S; an unbiased estimate of μ can be obtained as

$$\mu_p(S) = \frac{1}{n}\sum\nolimits_{i\in S} \frac{1}{\mathbb{E}_p[c_i]} x_i = \frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{c_i(S)}{\mathbb{E}_p[c_i]} x_i. \tag{2}$$

When sampling subsets uniformly (with or without replacement), we have $\mathbb{E}_p[c_i] = k/n$, recovering estimate (1). Alternatively, when $k = 1$ and the probability of sampling element i is given by $p_i = \|x_i\| / \sum_j \|x_j\|$, we recover the importance sampling estimate.

We seek the data-dependent distribution $p \in \mathcal{P}_{k,n}$ for which the estimate (2) achieves the lowest variance; our work thus focuses on solving the following problem:

$$\text{Find } p^* \in \operatorname{argmin}_{p\in\mathcal{P}_{n,k}} \mathbb{E}\big[\|\mu_p(S)\|^2\big].$$

Proposition 1. *Let $p \in \mathcal{P}_{k,n}$ be a distribution over multisets of fixed size k of $\{1, \dots, n\}$. Let $c_i(S)$ count the number of occurrences of i in a subset S. We write $\bar{c}_i = \mathbb{E}_{S\sim p}[c_i(S)]$ and $\bar{c}_{ij} = \mathbb{E}_{S\sim p}[c_i(S)c_j(S)]$. Let $\mu_p(S)$ be defined as in (2). Then, for $S \sim p$, $\mu_p(S)$ is an unbiased estimate of the mean of vectors x_i, and satisfies the equality*

$$\mathbb{E}\big[\|\mu_p(S)\|^2\big] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\bar{c}_{ij}}{\bar{c}_i\bar{c}_j} x_i^\top x_j. \tag{3}$$

Proposition 1 follows from a straightforward expansion of $\mathbb{E}[\|\mu_p(S)\|^2]$ using equation (2), and guides our search for optimal distribution to sample unbiased, low-variance estimates of the mean of a large set of high-dimensional points.

2.1 Characterizing a distribution by its 1st- and 2nd-order counts is NP-hard

One might be tempted to directly optimize the right-hand side of equality (3) to define an optimal distribution p. To do so, one must first identify the set of constraints on the first- and second-order counts $\bar{c}_i$ and $\bar{c}_{ij}$ so that they define a valid distribution $p \in \mathcal{P}_{n,k}$. Such constraints include, for example,

$$(i)\ \ \forall i, 0 \leq \bar{c}_i \leq k \qquad (ii)\ \ \sum\nolimits_i \bar{c}_i = k, \qquad (iii)\ \ \forall i, \sum\nolimits_{j \neq i} \bar{c}_{ij} = k\bar{c}_i - \bar{c}_{ii}.$$

Unfortunately, our first result is negative. Given $[\bar{c}_i]_{i\in[n]}$ and $[\bar{c}_{ij}]_{i,j\in[n]}$, determining if the $\bar{c}_i$ and $\bar{c}_{ij}$ correspond to counts of a distribution $p \in \mathcal{P}_{n,k}$ is NP-hard in the general case: even enumerating the constraints required to optimize (3) is, in practice, impossible.

We show this by reduction from the *unconstrained marginal decision problem* — deciding whether an assignment of first- and second-order counts correspond to a distribution over subsets of $[n]$ of *unconstrained* size — which was shown to be NP-hard in [21].

For the unconstrained problem, Sontag and Jaakkola [21] showed the NP-hardness result on distributions over sets sampled without replacement; for clarity purposes, we make the same assumptions in the proof below, although we note that distributions that sample with replacement can be represented by sampling without replacement over a larger set of duplicated items. When sampling without replacement, first order counts $\bar{c}_i$ correspond to the marginal probability $p(i \in S)$ of item i being sampled, and second-order counts $\bar{c}_{ij}$ correspond to the marginal probability$p(\{i, j\} \subseteq S)$ of $i \neq j$ being sampled together.

Lemma 2. *Let $\bar{c}_i, \bar{c}_{ij}$ be a set of possible counts on $[n]$. There exists a distribution p over subsets of any size of $[n]$ such that $p(i \in S) = c_i$ and $p(\{i, j\} \subseteq S) = \bar{c}_{ij}$ if and only if there exists a distribution $q \in \mathcal{Q}_{2n,n}$ that realizes the following marginals on $[2n]$, which we identify to $\{-n, \ldots, -1, 1, \ldots, n\}$:*

$$\begin{array}{ll} (i) & q(i \in S) = \bar{c}_i \\ (ii) & q(-i \in S) = 1 - \bar{c}_i \\ (iii) & q(\{i, j\} \subseteq S) = \bar{c}_{ij} \end{array} \qquad \begin{array}{ll} (iv) & q(\{i, -i\} \subseteq S) = 0 \\ (v) & q(\{i, -j\} \subseteq S) = \bar{c}_i - \bar{c}_{ij} \\ (vi) & q(\{-i, -j\} \subseteq S) = 1 - \bar{c}_i - \bar{c}_j + \bar{c}_{ij}. \end{array}$$

Corollary 3. *The following problem is NP-hard: given $[\bar{c}_i]_{i\in[2n]}$ and $[\bar{c}_{ij}]_{i,j\in[2n]}$, decide if there exists a distribution $p \in \mathcal{Q}_{2n,n}$ with first-order counts $\bar{c}_i$ and second-order counts $\bar{c}_{ij}$.*

Corollary 3 shows that directly optimizing the counts $\bar{c}_i$ and $\bar{c}_{ij}$, then seeking to find a distribution p with such first- and second-order counts, cannot be done without additional assumptions on the shape of the distribution. The rest of this paper focuses on finding distributions with unbiased, low-variance gradients when optimizing (3) under additional constraints that render the problem tractable.

2.2 Sampling sequentially with replacement

A natural simplifying assumption on the sampling distribution is that the k points are sampled sequentially and independently from each other; note that independence between

subsequent draws requires sampling with replacement. Hence, we begin our analysis by focusing on sampling k vectors sequentially from a fixed categorical distribution.

Proposition 4. *Let p be the distribution that samples a subset S of size k by sampling k points sequentially, each time from the categorical distribution such that point i is sampled with probability p_i. Then, setting $p_i = \|x_i\| / \sum_j \|x_j\|$ minimizes $\mathbb{E}\big[\|f_p(S)\|^2\big]$.*

Proposition 4 shows that importance sampling remains optimal in the batch setting ($k > 1$) when points are sampled via k sequential and *independent* draws. This is a crucial result for implementation purposes that rely on efficiency, as importance sampling only depends on the norm of the points x_i, and not on their pairwise inner products $x_i^\top x_j$.

For example, batch stochastic gradient descent can take advantage of methods that approximate the norm of the gradients, either by using stale gradients, or more prosaically by upper bounding the norm by the loss function's Lipschitz constant [16, 10], without relying on pairwise interactions between different training points.

2.3 Sampling without replacement

The analysis of batch sampling without replacement requires more careful analysis since two consecutive point samples are no longer independent events. Instead, as mentioned earlier, item counts in the without replacement setting now correspond to marginal probabilities: $\bar{c}_i = \bar{c}_{ii} = \Pr(i \in S)$, and, similarly, $\bar{c}_{ij} = \Pr(\{i, j\} \subseteq S)$. Hence, we reformulate our optimization problem in the following way:

$$\min_{p \in \mathcal{P}_{n,k}} f(p) \triangleq \frac{1}{n^2} \sum\nolimits_{i=1}^{n} \frac{1}{p_i} \|x_i\|^2 + \frac{1}{n^2} \sum\nolimits_{i \neq j} \frac{p_{ij}}{p_i p_j} x_i^\top x_j. \tag{4}$$

Equation (4) once again surfaces the importance sampling distribution: when the batch size is equal to one, the second order interactions disappear, and the optimal distribution over (single) points is the unique categorical distribution that minimizes $\sum_i \|x_i\|^2 / p_i$.

As identifying the set of constraints required for (4) is NP-hard, we must seek approximate solutions. A natural approach is to seek to only find the optimal values for the first order marginals $p_i = \Pr(i \in S)$, under a relaxed set of constraints. This discards the second term all-together, but significant insight can be gathered from just the marginal probabilities of singleton subsets. We therefore consider the problem:

$$\min_{0 \leq p_i \leq 1, \sum_{i=1}^{n} p_i = k} \sum\nolimits_{i=1}^{n} \|x_i\|^2 / p_i, \tag{5}$$

where the constraints sufficiently characterize marginal probabilities $p_1, \ldots, p_n \in [0, 1]$ corresponding to an expected final set size of k.

Proposition 5. *Let $x_1, \ldots, x_n$ be n vectors in $\mathbb{R}^d$; without loss of generality, we assume that the x_i are ordered such that $\|x_i\|_i$ forms an increasing sequence. Let $\kappa \leq k$ be the largest index such that $\|x_i\| \leq \frac{1}{\kappa} \sum_{j=1}^{n-k+\kappa} \|x_j\|$ for all $\in [n]$. The unique global minimum of (5) is achieved for $p_i = \kappa \frac{\|x_i\|}{\sum_{j=1}^{n-k+\kappa} \|x_j\|}$ for $i \in \{1, \ldots, n-k+\kappa\}$, and $p_i = 1$ otherwise.*

To summarize, the optimal first order-marginals are proportional to the norm $\|x_i\|$ (as in importance sampling), unless there exists at least one point i such that $\|x_i\| / \sum_j \|x_j\| > 1/k$.

Any such point is assigned a marginal sampling probability $p_i = 1$, guaranteeing that it will always be sampled. In the case where $k \ll n$, such points will occur only when the distribution of vector norms $\|x_i\|$ is highly skewed.

2.4 Learning a parametric distribution to sample without replacement

Importance sampling has the drawback of ignoring the potentially significant variance information contained in the second order marginals. As a result, importance sampling has no control over redundancy between items. To improve upon this defect we propose a simple alternative approach by learning the sample from a parameterized class $\mathcal{F} = \{p(\cdot;\theta) \mid \theta\}$ of distributions over subsets of size k, we can instead to find the best parameterization,

$$\theta^* \in \operatorname*{argmin}_{\theta} \mathbb{E}_{p(\cdot;\theta)}\Big[\|\mu_{p(\cdot;\theta)}(S)\|^2\|\Big]. \tag{6}$$

Among distributions over fixed-sized subsets of a ground set, k-determinantal point processes [12] (k-DPPs) are among the best known. DPPs and k-DPPs are known to be distributions that favor subsets of high-quality yet *diverse* items [13]; thus, DPPs have found many applications to variance reduction problems, including experimental design [4] and minibatch sampling [23]. Under a k-DPP parameterized by a kernel matrix $L \in \mathbb{R}^{n\times n} \succeq 0$, any subset of size k has probability

$$P(S) \propto \det(L_S),$$

where $L_S = [L_{ij}]_{i,j\in S}$ is the principal submatrix of L indexed by items in S. k-DPPs are particularly well suited to being learned to optimize sampling variance, as their first- and second-order marginals admit a closed-form expression: under a k-DPP with kernel $L \succeq 0$,

$$p_i = \frac{Z_{k-1}^{\{i\}}}{Z_k} L_{ii} \qquad\qquad p_{ij} = \frac{Z_{k-2}^{\{i,j\}} Z_k}{Z_{k-1}^{\{i\}} Z_{k-1}^{\{j\}}} \left(L_{ii}L_{jj} - L_{ij}^2\right),$$

where for any subset A, the normalization coefficient $Z_{k-|A|}^A$ is defined as [13, Section 5.2.3]

$$Z_{k-|A|}^A = E_{k-|A|}\Big(\Big(\big[(L + I_{\bar{A}})^{-1}\big]_{\bar{A}}\Big)^{-1} - I\Big),$$

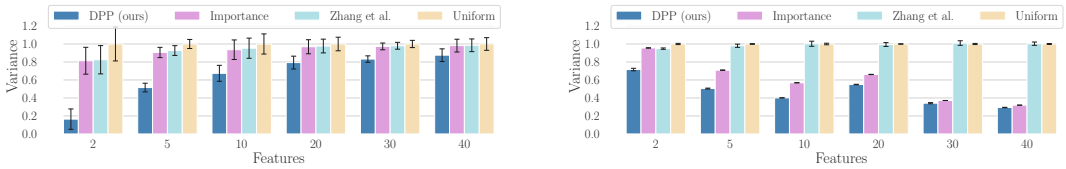
where E_k is the kth elementary symmetric polynomial on the eigenvalues of matrices, and we denote by $\bar{A}$ the complement of A in $[n]$.

3. Experiments

We verify our proposed variance reduction method by learning a determinantal point process (DPP) to sample a batch. We begin by verifying empirically that when minimizing (3), the resulting samples achieve lower variance. Given $\mathcal{Y} = \{x_1, \ldots, x_n\}$ a ground set of n vectors in $\mathbb{R}^d$, we evaluate against the following baselines:

- UNIF: the uniform distribution over subsets of size k
- IS: sampling k times without replacement from the importance sampling distribution
- DPP-ZHANG: The k-DPP with radial kernel $L_{ij} = e^{-\|x_i - x_j\|^2}$, as indicated in [23].

As seen in Figures 1a and 1b, the DPP that minimizes the cost function (3) by optimizing its first and second order marginals consistently achieves a significantly smaller variance than other baseline methods.



(a) Vectors x_i are obtained by drawing n times from a mixture of Gaussians

(b) Vectors x_i are per-element gradients of a logistic regression model after 100 training iterations.

Figure 1: Batch variance under different distributions, as a function of vector dimension d. Learned DPP that explicitly optimize for their first and second-order marginals outperform all other methods, consistently across data distribution types and dimensionality.

We also investigate the optimization of a logistic model using different batch sampling methods.[2] As comparing loss curves of learning methods is notoriously difficult, for each stochastic gradient step we select the optimal learning rate η by line-search with increments of 0.001. This is important, as it has been observed that batches with lower variances can accommodate larger learning rates [9]. Fig. 2 shows the learned DPP method converges in around 4 times fewer iterations than importance sampling, and significantly faster than uniform sampling, and the DPP based method of Zhang et al. [23].

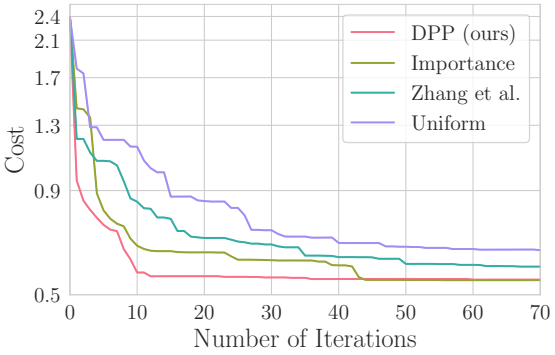


Figure 2: Training loss of a logistic model trained using batched-SGD, using different batch samplers.

4. Conclusion and future work

Approximating the mean of sets of vectors via subset-selection is a crucial machine learning task; most importantly, it arises in batched stochastic estimates of full gradients when training on large datasets. We show that a distribution over batches of points that achieves unbiased estimates of minimal variance can be characterized by marginal element counts.

Although it is in general NP-hard to use optimal marginal counts to define a distribution, when sampling points sequentially with replacement, this characterization can be used to prove that importance sampling remains optimal. Without replacement, importance sampling optimality is recovered under strong simplifying assumptions. Without those assumptions, we use the marginal characterization of optimality to *learn* a parametric distribution over batches. Experimentally, when learning a DPP in such a way, we obtain improved low-variance estimates over various vector sets; in a downstream SGD experiment, we show that learning approximately optimal DPPs achieves a faster convergence rate.

In this paper, we made the standard assumption that we have access to oracles that provide vector norms and inner products; extensions of this work will require using inexpensive approximations in lieu of these oracles, as has been done for example in [24, 15, 10]. Furthermore, although DPPs are a natural choice of distribution over fixed sized sets, other choices are possible; in particular, dual volume sampling distributions [2, 14, 17] also provide closed-form marginals and enjoy similar negative-depndence properties as DPPs.

2. This serves as a proof of concept, as learning a logistic model is a convex optimization problem.

References

[1] Guillaume Alain, Alex Lamb, Chinnadhurai Sankar, Aaron Courville, and Yoshua Bengio. Variance reduction in SGD by distributed importance sampling. *arXiv preprint arXiv:1511.06481*, 2015.

[2] Haim Avron and Christos Boutsidis. Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications*, 34(4), 2013.

[3] Dominik Csiba and Peter Richtárik. Importance sampling for minibatches. *The Journal of Machine Learning Research*, 19(1), 2018.

[4] Michal Derezinski, Feynman Liang, and Michael W. Mahoney. Bayesian experimental design using regularized determinantal point processes. *CoRR*, abs/1906.04133, 2019.

[5] G. Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2), 1952.

[6] Tianfan Fu and Zhihua Zhang. CPSG-MCMC: Clustering-Based Preprocessing method for Stochastic Gradient MCMC. volume 54 of *Proceedings of Machine Learning Research*, 2017.

[7] Yingjie Gu and Zhong Jin. Neighborhood preserving d-optimal design for active learning and its application to terrain classification. *Neural Computing and Applications*, 23 (7-8), 2013.

[8] Xiaofei He. Laplacian regularized d-optimal design for active learning and its application to image retrieval. *Trans. Img. Proc.*, 19(1), January 2010.

[9] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, 2013.

[10] A. Katharopoulos and F. Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[11] J. Kiefer. Optimal design: Variation in structure and performance under change of criterion. *Biometrika*, 62(2), 08 1975.

[12] Alex Kulesza and Ben Taskar. K-dpps: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11. Omnipress, 2011.

[13] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. 2012.

[14] Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Polynomial time algorithms for dual volume sampling. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017.

[15] Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *ArXiv*, abs/1511.06343, 2015.

[16] Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. 2016.

[17] Zelda E. Mariet and Suvrit Sra. Elementary symmetric polynomials for optimal experimental design. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017.

[18] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014.

[19] Friedrich Pukelsheim. *Optimal Design of Experiments (Classics in Applied Mathematics, 50)*. Society for Industrial and Applied Mathematics, USA, 2006.

[20] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, 2016.

[21] David Sontag and Tommi Jaakkola. On iteratively constraining the marginal polytope for approximate inference and map, 2007.

[22] Chong Wang, Xi Chen, Alexander J Smola, and Eric P Xing. Variance reduction for stochastic gradient optimization. In *Advances in Neural Information Processing Systems*, 2013.

[23] Cheng Zhang, Hedvig Kjellstrom, and Stephan Mandt. Determinantal point processes for mini-batch diversification. *arXiv preprint arXiv:1705.00607*, 2017.

[24] Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37. PMLR, 2015.

Appendix A. Proofs

A.1 NP-hardness

Lemma 2. Let $\bar{c}_i, \bar{c}_{ij}$ be a set of possible counts on $[n]$. There exists a distribution p over subsets of any size of $[n]$ such that $p(i \in S) = c_i$ and $p(\{i,j\} \subseteq S) = \bar{c}_{ij}$ if and only if there exists a distribution $q \in \mathcal{Q}_{2n,n}$ that realizes the following marginals on $[2n]$, which we identify to $\{-n, \dots, -1, 1, \dots, n\}$:

$$
\begin{array}{ll}
(i) & q(i \in S) = \bar{c}_i \\
(ii) & q(-i \in S) = 1 - \bar{c}_i \\
(iii) & q(\{i,j\} \subseteq S) = \bar{c}_{ij}
\end{array}
\qquad
\begin{array}{ll}
(iv) & q(\{i,-i\} \subseteq S) = 0 \\
(v) & q(\{i,-j\} \subseteq S) = \bar{c}_i - \bar{c}_{ij} \\
(vi) & q(\{-i,-j\} \subseteq S) = 1 - \bar{c}_i - \bar{c}_j + \bar{c}_{ij}.
\end{array}
$$

Proof. We identify each subset of S of $[n]$ with a k-subset $\hat{S}$ of $[2n] \equiv \{-n, \dots, -1, 1, \dots, n\}$:

$$\hat{S} = S \cup \{-i \mid i \notin S\}.$$

This induces a bijection between distributions $q \in \mathcal{P}_{2n,n}$ that realize marginals (i) – (vi) and distributions $p \in \mathcal{Q}_n$ that realize $\bar{c}_{ij}$; for $S \subseteq [n]$, $q(\hat{S}) = q(S)$; q is zero on all sets of size $n' \neq n$.

If there exists a distribution $q \in \mathcal{P}_{2n,n}$ realizing the $\bar{c}_{ij}$, then the existence of $P \in \mathcal{P}_n$ is trivial by (iii). Now, assume there exists a distribution $p \in \mathcal{P}_n$ realizing the counts $\bar{c}_{ij}$. The distribution q defined by $q(\hat{S}) = p(S)$ realizes marginals q_{ij} as defined by (i) – (vi).

$$
\begin{aligned}
q_{ij} &= \sum_{\hat{S} \ni i,j} q(\hat{S}) = \sum_{S \ni i,j} p(S) = \bar{c}_{ij} \\
q_{i,-i} &= \sum_{\hat{S} \ni i,-i} q(\hat{S}) = \sum_{S \ni i, S \not\ni i} p(S) = 0 \\
q_{i,-j} &= \sum_{\hat{S} \ni i,-j} q(\hat{S}) = \sum_{\substack{S \ni i \\ S \not\ni j}} p(S) = \bar{c}_i - \bar{c}_{ij} \\
q_{-i,-j} &= \sum_{\hat{S} \ni -i,-j} q(\hat{S}) = \sum_{S \not\ni i,j} p(S) = 1 - \bar{c}_i - \bar{c}_j + \bar{c}_{ij}
\end{aligned}
$$

■

Corollary 3. The following problem is NP-hard: given $[\bar{c}_i]_{i \in [2n]}$ and $[\bar{c}_{ij}]_{i,j \in [2n]}$, decide if there exists a distribution $p \in \mathcal{Q}_{2n,n}$ with first-order counts $\bar{c}_i$ and second-order counts $\bar{c}_{ij}$.

Proof. By Lemma 2 we reduce the problem of deciding whether an assignment of first- and second-order counts corresponds to a distribution over subsets of $[n]$ to deciding whether a corresponding distribution exists over subsets of size n from a ground set of $2n$ items. As the first problem is NP-hard, its reduction is also NP-hard. ■

A.2 Optimality of importance sampling

Proposition 4. Let p be the distribution that samples a subset S of size k by sampling k points sequentially, each time from the categorical distribution such that point i is sampled with probability p_i. Then, setting $p_i = \|x_i\| / \sum_j \|x_j\|$ minimizes $\mathbb{E}\big[\|f_p(S)\|^2\big]$.

This result depends upon the following equalities:

Lemma 6. *Under the conditions of Prop. 4,*

$$\bar{c}_{ii} = kp_i(1 - p_i + kp_i) \qquad \text{and} \qquad \bar{c}_{ij} = k(k-1)p_i p_j.$$

Proof. We begin by obtaining two equalities similar to the first and second moments of the binomial distribution. Let $p, q \in [0, 1]$, and let $f(t)$ be the generalized moment-generating function for the binomial distribution:

$$f(t) = \sum_{i=0}^{n} e^{it} \binom{i}{n} p^i q^{n-i} = (pe^t + q)^n.$$

By differentiating this equality on both sides and evaluating it at $t = 0$, we obtain

$$\sum_{i=0}^{n} \binom{n}{i} i p^i q^{n-i} = np(p+q)^{n-1}. \tag{7}$$

Similarly, by another differentiation, we obtain

$$\sum_{i=0}^{n} \binom{n}{i} i^2 p^i q^{n-i} = np(p+q)^{n-2}(np+q). \tag{8}$$

Setting $q = 1 - p$, we recover the first and second (non-central) moments of the binomial distribution. With (7) and (8) in hand, we can now prove Lemma 6.

Let p be the distribution over subsets of size k that samples k times, independently, from the categorical distribution over $[n]$ parameterized by $p_1, \ldots, p_n \in [0, 1]$ (where p_i indicates the probability of sampling point i at any step).

By definition, we have

$$\begin{aligned}
\bar{c}_{ii} &= \mathbb{E}_{S\sim p}[c_i^2] \\
&= \sum_{m=0}^{k} \Pr(c_i = m) m^2 \\
&= \sum_{m=0}^{k} \binom{k}{m} p_i^m (1 - p_i)^{k-m} m^2 \\
&= kp(kp_i + 1 - p_i),
\end{aligned}$$

where the last equality follows from (8).

Similarly, for $i \neq j$, we have

$$\begin{aligned}
\bar{c}_{ij} =& \mathbb{E}_{S\sim p}[c_i c_j] \\
&= \sum_{m=0}^{k} m \binom{k}{m} p_i^m \times \sum_{m'=0}^{k-m} \binom{k-m}{m'} m' p_j^{m'} (1 - p_i - p_j)^{k-m-m'} \\
&\overset{(a)}{=} \sum_{m=0}^{k} m \binom{k}{m} p_i^m \left[(k-m) p_j (1 - p_i)^{k-m-1}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{kp_j}{1-p_i} \sum_{m=0}^{k} m \binom{k}{m} p_i^m (1-p_i)^{k-m} - \frac{p_j}{1-p_i} \sum_{m=0}^{k} m^2 \binom{k}{m} p_i^m (1-p_i)^{k-m} \\
&\overset{(b)}{=} k^2 \frac{p_i p_j}{1-p_i} - \frac{p_j}{1-p_i} \big[kp_i(kp_i + 1 - p_i) \big] \\
&= k(k-1) p_i p_j,
\end{aligned}
$$

where equality (a) follows from (7) and (b) follows from (7) and (8), concluding the proof. ■

Proof (Prop. 4). After replacing the counts $\bar{c}_{ij}$ in equation (3) with the identities in Lemma 6, it appears that the only term of $\mathbb{E}\big[\|\mu_p(S)\|^2\big]$ that depends on the p_i is of the form

$$
\frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{kp_i} \|x_i\|^2,
$$

which is known to be minimized by importance sampling, *i.e.*, $p_i \propto \|x_i\|$. ■

A.3 Importance sampling without replacement

$$
\min_{p \in \mathcal{P}_{n,k}} T_1(p) = \min_{0 \le p_i \le 1, \sum_{i=1}^n p_i = k} \sum_{i=1}^{n} \frac{1}{p_i} \|x_i\|^2. \tag{9}
$$

Note that this objective function is a sum of strictly convex functions, and hence strictly convex, and the feasible set is also convex. Therefore there exists a unique global minimum. We shall assume without loss of generality for the rest of this section that the gradients are ordered so that $\{\|x_i\|\}_{i=1}^n$ is a non-decreasing sequence.

Proposition 7. *Let $\kappa \in [k]$ be the largest element of $[k]$ such that $\|x_i\| \le \frac{1}{\kappa} \sum_{j=1}^{n-k+\kappa} \|x_j\|$ for all $j \in [n]$. The unique global minimum of (9) is $p_i = 1$ for all $i \in [n] \setminus [n-k+\kappa]$ and $p_i = \kappa \frac{\|x_i\|}{\sum_{j=1}^{n-k+\kappa} \|x_j\|}$ for $i \in [n-k+\kappa]$.*

Note that in the case $k \ll n$, it is likely that $\|x_j\| \le \frac{1}{k} \sum_{i=1}^{n} \|x_i\|$ for all $j \in [n]$. In this case the optimal solution is simply

$$
p_i = k \frac{\|x_i\|}{\sum_{j=1}^{n} \|x_j\|}
$$

for all $i \in [n]$. This proposition follows as a corollary of the following two lemmas.

Lemma 8. *Let $\|x_i\| \le \frac{1}{k} \sum_{j=1}^{n} \|x_j\|$ for all $i \in [n]$. Then the optimal solution of*

$$
\min_{0 \le p_i \le 1, \sum_{i=1}^n p_i = k} \sum_{i=1}^{n} \frac{1}{p_i} \|x_i\|^2 \tag{10}
$$

is $p_i = k \frac{\|x_i\|}{\sum_{j=1}^{n} \|x_j\|}$ for all $i \in [n]$.

Proof. We note first that the claimed solution is clearly feasible. Now we will show that the given p_i's solve

$$\min_{0\le p_i, \sum_{i=1}^n p_i = k} \sum_{i=1}^n \frac{1}{p_i}\|x_i\|^2. \tag{11}$$

The Lagrangian for this problem is

$$\mathcal{L} = \sum_{i=1}^n \frac{1}{p_i}\|x_i\|^2 + \mu\left(\sum_{i=1}^n p_i - k\right) - \sum_{i=1}^n \lambda_i p_i$$

At optimality, the KKT stationarity conditions imply $0 = \partial_{p_i}\mathcal{L} = -(\frac{\|x_i\|}{p_i})^2 + \mu - \lambda_i$ and feasibility implies $0 = \partial_\mu \mathcal{L} = \sum_{i=1}^n p_i - k$. Furthermore, the KKT complementary slackness condition implies that $\lambda_i p_i = 0$. Since the objective cannot be optimal if $p_i = 0$ for any i, we conclude that $\lambda_i = 0$ for all i. Inserting this result into the stationarity condition for p_i we find that $(\frac{\|x_i\|}{p_i})^2 = \mu$, finally yielding

$$p_i = \frac{\|x_i\|}{\sqrt{\mu}}$$

Finally, combining this with the stationarity condition for μ, we conclude that

$$p_i = k\frac{\|x_i\|}{\sum_{j=1}^n \|x_j\|}$$

is an optimal solution to problem 11. Since this solution also satisfies $p_i \le 1$ for all i it must also be an optimal solution to problem 10. ■

Lemma 9. *Suppose that* $\|x_n\| > \frac{1}{k}\sum_{j=1}^n \|x_j\|$. *Then,*

$$\min_{0\le p_i\le 1, \sum_{i=1}^n p_i = k} \sum_{i=1}^n \frac{1}{p_i}\|x_i\|^2 = \|x_n\|^2 + \min_{0\le p_i\le 1, \sum_{i=1}^{n-1} p_i = k-1} \sum_{i=1}^{n-1} \frac{1}{p_i}\|x_i\|^2$$

Proof. Let $\{p_i^*\}_{i\in[n-1]}$ be an optimal solution to the right hand problem. Then $p_n = 1$ and $p_i = p_i^*$ for $i \in [n-1]$ defines a feasible solution to the left hand problem with the same objective value as the right hand problem. This implies that LHS $\le$ RHS. To prove the lemma it therefore suffices to show that any optimal solution to the LHS has $p_n = 1$. The Lagrangian for the LHS problem is

$$\mathcal{L} = \sum_{i=1}^n \frac{1}{p_i}\|x_i\|^2 + \mu\left(\sum_{i=1}^n p_i - k\right) - \sum_{i=1}^n \lambda_i p_i + \sum_{i=1}^n \eta_i(p_i - 1)$$

The KKT stationarity conditions imply $0 = \partial_{p_i}\mathcal{L} = -(\frac{\|x_i\|}{p_i})^2 + \mu - \lambda_i + \eta_i$ and feasibility implies $0 = \partial_\mu \mathcal{L} = \sum_{i=1}^n p_i - k$. Furthermore, the KKT complementary slackness condition implies that $\lambda_i p_i = 0$ and $\eta_i(p_i - 1) = 0$. Again we know that $p_i > 0$ for any optimal solution, so $\lambda_i = 0$ for all i. Furthermore, suppose $p_i < 1$ for all i. Then $\eta_i = 0$ and the stationary condition and feasibility condition imply that $p_i = k\frac{\|x_i\|}{\sum_{j=1}^n \|x_j\|}$ for all i. However

since $\|x_n\| > \frac{1}{k}\sum_{j=1}^{n}\|x_j\|$ this solution is not feasible. Therefore for an optimal solution $\{p_i\}_{i=1}^n$ there must exist an i^* such that $p_{i^*} = 1$. If $i^* = n$ then we are done, so suppose $i^* \neq n$. We shall now construct a feasible solution $\tilde{p}$ which is at least as good as p and for which $\tilde{p}_n = 1$. Define $\tilde{p}_n = p_{i^*} = 1$, $\tilde{p}_{i^*} = p_n$ and $\tilde{p}_i = p_i$ for all $i \neq i^*, n$. Since p is feasible, so too is $\tilde{p}$. Furthermore,

$$\begin{aligned}\sum_{i=1}^{n}\frac{1}{\tilde{p}_i}\|x_i\|^2 &= \sum_{i\neq i^*,n}\frac{1}{\tilde{p}_i}\|x_i\|^2 + \frac{1}{\tilde{p}_n}\|x_{i^*}\|^2 + \frac{1}{\tilde{p}_{i^*}}\|x_n\|^2 \\ &\leq \sum_{i\neq i^*,n}\frac{1}{p_i}\|x_i\|^2 + \frac{1}{p_{i^*}}\|x_{i^*}\|^2 + \frac{1}{p_n}\|x_n\|^2 \\ &= \sum_{i=1}^{n}\frac{1}{p_i}\|x_i\|^2\end{aligned}$$

where the inequality holds since $p_{i^*} \geq p_n$ and $\|x_n\| \geq \|x_{i^*}\|$ by assumption. ■

Proof of Proposition 7. We describe an iterative procedure that yields the claimed solution upon completion.

1. Initialize $\kappa \leftarrow k$.
2. Check if the hypothesis of Lemma 8 hold for $\{\|x_i\|\}_{i=1}^{n-k+\kappa}$. If they do, set $p_i = \kappa\frac{\|x_i\|}{\sum_{j=1}^{n-k+\kappa}\|x_j\|}$ for $i \in [n-k+\kappa]$, and return $\{p_i\}_{i=1}^n$ and κ.
3. Otherwise, by Lemma 9, set $p_{n-k+\kappa} = 1$, then update $\kappa \leftarrow \kappa - 1$, and loop back to step 2.

Note that this process must terminate, at the latest, when $\kappa = 1$. By repeated applications of Lemmas 8 and 9, the returned solution is optimal. Further, by consulting the definition of κ in Proposition 7, one observes that it coincides with the value of κ returned by this iterative procedure. ■

Appendix B. Experimental details

Optimizing Dpp We minimize the variance objective as a function of Φ where $L = \Phi^\top\Phi$ is the Dpp kernel. We initialize Φ such that the kernel is the kernel $L_{ij} = e^{-\|x_i - x_j\|^2}$ proposed by Zhang et al. [23]. To optimize Φ we use Adam with learning rate 0.2, and $(\beta_1, \beta_2) = (0.9, 0.999)$. We train for 150 steps, reducing the learning rate by a factor of 0.3 after 100 steps.

Synthetic Data For figure 2 we sample 1000 data points from a mixture of four 10-dimensional Gaussians, where two clusters correspond to class +1 and the other two clusters correspond to class −1.

Logistic Regression Optimization Batches of size 5 are selected from a uniformly at random sub-sampled set of 30 data points. The learning rate is decided at each iteration by line search with increment 0.001. For reproducibiliy we initialize the random seed of PyTorch and Numpy to be 325.