| **CPSC 486/586: Probabilistic Machine Learning** | *Out: January 18, 2023* |
|---|---|
| Problem Set 1 | |
| *Instructor: Andre Wibisono* | *Due: February 1, 2023* |

(P1) Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be the quadratic function: $f(x) = \frac{1}{2}\|x\|^2$.

(a) Write down the gradient descent algorithm with step size $\eta > 0$:

$$(\mathsf{GD}_f): \qquad x_{k+1} = x_k - \eta \nabla f(x_k).$$

Solve the recursion and determine how fast $x_k$ converges to its limit $x^*$ as $k \to \infty$. What is $x^*$? For which step size $\eta$ does the conclusion hold?

**Solution:** Note that $\nabla f(x) = x$. The gradient descent update is:

$$x_{k+1} = x_k - \eta x_k = (1-\eta)x_k.$$

Starting from $k = 0$:

$$\begin{aligned} x_1 &= (1-\eta)x_0 \\ x_2 &= (1-\eta)x_1 = (1-\eta)^2 x_0 \\ &\vdots \\ x_k &= (1-\eta)^k x_0. \end{aligned}$$

For $0 < \eta \le 1$, the limit is $\lim_{k\to\infty}(1-\eta)^k x_0 = \mathbf{0} = x^*$, otherwise it diverges. One can also see through first-order optimality conditions for $f$ that $x^* = \mathbf{0}$.

(b) Write down the proximal gradient method with step size $\eta > 0$:

$$(\mathsf{PG}_f): \qquad x_{k+1} = \arg\min_{x\in\mathbb{R}^d} f(x) + \frac{1}{2\eta}\|x - x_k\|^2.$$

Solve the recursion and determine how fast $x_k$ converges to its limit $x^*$ as $k \to \infty$. For which step size $\eta$ does the conclusion hold?

**Solution:** Note that $g(x) = f(x) + \frac{1}{2\eta}\|x - x_k\|^2$ is convex in $x$, so we can solve the minimization problem using the first-order optimality condition. The minimizer is $x^*$ such that

$$\nabla g(x^*) = 0 \iff x^* + \frac{1}{\eta}x^* - \frac{1}{\eta}x_k = 0 \iff \left(1 + \frac{1}{\eta}\right)x^* = \frac{1}{\eta}x_k \iff x^* = \frac{1}{1+\eta}x_k.$$

Starting from $k = 0$:

$$\begin{aligned} x_1 &= \arg\min_{x \in \mathbb{R}^d} \left\{ \frac{1}{2}\|x\|^2 + \frac{1}{2\eta}\|x - x_0\|^2 \right\} = \frac{x_0}{1+\eta} \\ x_2 &= \frac{x_1}{1+\eta} = \frac{x_0}{(1+\eta)^2} \\ &\vdots \\ x_k &= \frac{x_0}{(1+\eta)^k}. \end{aligned}$$

For any $\eta > 0$, $\lim_{k\to\infty} x_k = \lim_{k\to\infty} \frac{x_0}{(1+\eta)^k} = \mathbf{0}$. Note there is no restriction on how large $\eta$ can be.

(c) Write down the solution to the gradient flow dynamics:

$$(\mathsf{GF}_f): \qquad \dot{X}_t = -\nabla f(X_t).$$

Write $X_t$ in terms of $X_0$ and $t$, and determine how fast $X_t$ converges to its limit $x^*$ as $t \to \infty$. Suppose $t = \eta k$. How does your answer compare to part (a) and (b)?

**Solution:** The gradient flow dynamics becomes

$$\dot{X}_t = -X_t.$$

This is a linear ODE which we can solve explicitly. Consider the time derivative of $Y_t = e^t X_t$:

$$\dot{Y}_t = e^t X_t + e^t \dot{X}_t = e^t(\dot{X}_t + X_t) = 0.$$

This means $Y_t$ is a constant, so for all $t \geq 0$:

$$e^t X_t = Y_t = Y_0 = X_0.$$

Thus, the solution is for all $t \geq 0$:

$$X_t = e^{-t} X_0$$

Note $X_t$ converges to $x^* = 0$ exponentially fast as $t \to \infty$:

$$\|X_t - x^*\| = \|X_t\| = e^{-t}\|X_0\|.$$

(The right hand side above is a multiple of $e^{-t}$ which decreases to 0 exponentially fast.) Let us compare this convergence of GF with the convergence of GD and PG above. Suppose we choose step size $\eta > 0$ to discretize time, so each discrete-time iteration corresponds to an elapse of $\eta$ time in continuous time. Thus, $k$ iterations in discrete time corresponds to time $t = k\eta$ in continuous time.

For GD, note the convergence rate is:

$$\|x_k\| = (1-\eta)^k \|x_0\|.$$

Note for small $0 < \eta \ll 1$, $1 - \eta \approx e^{-\eta}$ (since $e^{-\eta} = 1 - \eta + \frac{1}{2}\eta^2 + \cdots$). Thus, the convergence rate of GD matches the continuous-time rate of GF:

$$(1-\eta)^k \approx e^{-\eta k} = e^{-t}.$$

Similarly, for PG, note the convergence rate is:

$$\|x_k\| = \frac{\|x_0\|}{(1+\eta)^k}.$$

Note for small $0 < \eta \ll 1$, $\frac{1}{1+\eta} \approx e^{-\eta}$ (since $e^{\eta} = 1+\eta+\frac{1}{2}\eta^2+\cdots$). Thus, the convergence rate of PG matches the continuous-time rate of GF:

$$\frac{1}{(1+\eta)^k} \approx e^{-\eta k} = e^{-t}.$$

(You can also note that $1 - \eta \le e^{-\eta} \le \frac{1}{1+\eta}$ for $\eta > 0$, which means $\mathsf{PG}_f$ is slower than $\mathsf{GF}_f$, which is slower than $\mathsf{GD}_f$. This is for the case when $f(x) = \frac{1}{2}\|x\|^2$. Can you show this holds more generally, e.g. for convex $f$?)

(P2) Suppose we want to minimize $h\colon \mathbb{R} \to \mathbb{R}$ given by $h(x) = f(x) + g(x)$ where $f(x) = \frac{1}{2}(x-1)^2$ and $g(x) = \frac{1}{2}(x+1)^2$. Suppose we can only use the gradient descent $\mathsf{GD}_f, \mathsf{GD}_g$ or proximal gradient $\mathsf{PG}_f, \mathsf{PG}_g$ for $f$ and $g$ separately (but we cannot do $\mathsf{GD}_{f+g}$ or $\mathsf{PG}_{f+g}$).

(a) For each algorithm below, write down its recursion explicitly, and determine its limit $x^*_\eta = \lim_{k\to\infty} x_k$.

**Solution:** Note that the algorithm updates are, explicitly:

$$\begin{aligned}
\mathsf{GD}_f(x_k) &= x_k - \eta(x_k - 1) = (1-\eta)x_k + \eta \\
\mathsf{GD}_g(x_k) &= x_k - \eta(x_k + 1) = (1-\eta)x_k - \eta \\
\mathsf{PG}_f(x_k) &= \arg\min_x \left\{ f(x) + \frac{1}{2\eta}(x - x_k)^2 \right\} = \frac{1}{1+\eta}x_k + \frac{\eta}{1+\eta} \\
\mathsf{PG}_g(x_k) &= \arg\min_x \left\{ g(x) + \frac{1}{2\eta}x_k(x - x_k)^2 \right\} = \frac{1}{1+\eta}x_k - \frac{\eta}{1+\eta}
\end{aligned}$$

i. $x_{k+1} = \mathsf{GD}_g \circ \mathsf{GD}_f(x_k)$

**Solution:** For any $k$,

$$\begin{aligned}
x_{k+1} &= \mathsf{GD}_g \circ \mathsf{GD}_f(x_k) \\
&= \mathsf{GD}_g((1-\eta)x_k + \eta) \\
&= (1-\eta)\left[(1-\eta)x_k + \eta\right] - \eta \\
&= (1-\eta)^2 x_k - \eta^2
\end{aligned}$$

To find its limit $x^*_\eta$, we can proceed in two ways. We can find the explicit solution to $x_k$, and take the limit $k \to \infty$ (see below). We can also find $x^*_\eta$ by solving for the point which does not change under the update: $x^*_\eta = \mathsf{GD}_g \circ \mathsf{GD}_f(x^*_\eta)$, or:

$$x^*_\eta = (1-\eta)^2 x^*_\eta - \eta^2.$$

This gives $x^*_\eta = \frac{-\eta}{2-\eta}$, which agrees with the calculation below.

To find the explicit solution of $x_k$, we can unroll the recursion:

$$\begin{aligned}
x_1 &= (1-\eta)^2 x_0 - \eta^2 \\
x_2 &= (1-\eta)^2 x_1 - \eta^2 = (1-\eta)^4 - \eta^2\left[(1-\eta)^2 + 1\right] \\
x_3 &= (1-\eta)^2 x_2 - \eta^2 = (1-\eta)^6 - \eta^2\left[(1-\eta)^4 + (1-\eta)^2 + 1\right] \\
&\vdots \\
x_k &= (1-\eta)^{2k} - \eta^2 \sum_{i=0}^{k-1} (1-\eta)^{2i}
\end{aligned}$$

As $k \to \infty$, $(1-\eta)^{2k} \to 0$. The second term is an infinite series with

$$\sum_{i=0}^{\infty} (1-\eta)^{2i} = \frac{1}{1-(1-\eta)^2} = \frac{1}{\eta(2-\eta)} \Rightarrow -\eta^2 \sum_{i=0}^{\infty} (1-\eta)^{2i} = \frac{-\eta}{2-\eta}.$$

Thus we have that $x^*_\eta = \lim_{k\to\infty} x_k = \frac{-\eta}{2-\eta}$.

ii. $x_{k+1} = \mathsf{GD}_g \circ \mathsf{PG}_f(x_k)$

**Solution:** For any $k$,

$$\begin{aligned}
x_{k+1} &= \mathsf{GD}_g \circ \mathsf{PG}_f(x_k) \\
&= \mathsf{GD}_g\left(\frac{1}{1+\eta}x_k + \frac{\eta}{1+\eta}\right)
\end{aligned}$$

$$
\begin{aligned}
&= (1-\eta)\left(\frac{1}{1+\eta}x_k + \frac{\eta}{1+\eta}\right) - \eta \\
&= \frac{1-\eta}{1+\eta}x_k
\end{aligned}
$$

Then $x_k = \left(\frac{1-\eta}{1+\eta}\right)^k x_0$, and we have $\lim_{k\to\infty} x_k = 0$ since $0 < \frac{1-\eta}{1+\eta} < 1$ for $0 < \eta < 1$.

iii. $x_{k+1} = \mathsf{PG}_g \circ \mathsf{GD}_f(x_k)$

**Solution:** For any $k$,

$$
\begin{aligned}
x_{k+1} &= \mathsf{PG}_g \circ \mathsf{GD}_f(x_k) \\
&= \mathsf{PG}_g\left((1-\eta)x_k + \eta\right) \\
&= \frac{1}{1+\eta}\left((1-\eta)x_k + \eta\right) - \frac{\eta}{1+\eta} \\
&= \frac{1-\eta}{1+\eta}x_k
\end{aligned}
$$

Thus, we have $x_k = \left(\frac{1-\eta}{1+\eta}\right)^k x_0$ and $\lim_{k\to\infty} x_k = 0$ since $0 < \frac{1-\eta}{1+\eta} < 1$ for $0 < \eta < 1$.

iv. $x_{k+1} = \mathsf{PG}_g \circ \mathsf{PG}_f(x_k)$

**Solution:**

$$
\begin{aligned}
x_{k+1} &= \mathsf{PG}_g \circ \mathsf{PG}_f(x_k) \\
&= \mathsf{PG}_g\left(\frac{1}{1+\eta}x_k + \frac{\eta}{1+\eta}\right) \\
&= \frac{1}{1+\eta}\left(\frac{1}{1+\eta}x_k + \frac{\eta}{1+\eta}\right) - \frac{\eta}{1+\eta} \\
&= \left(\frac{1}{1+\eta}\right)^2 x_k - \eta^2\left(\frac{1}{1+\eta}\right)^2.
\end{aligned}
$$

The explicit solution is

$$
x_k = \left(\frac{1}{1+\eta}\right)^{2k} x_0 - \eta^2 \sum_{i=1}^{k}\left(\frac{1}{1+\eta}\right)^{2i}
$$

As $k \to \infty$, the first term $\left(\frac{1}{1+\eta}\right)^{2k} \to 0$. The second term is an infinite series with

$$
\sum_{i=1}^{k}\left(\frac{1}{1+\eta}\right)^{2i} = \frac{\left(\frac{1}{1+\eta}\right)^2}{1-\left(\frac{1}{1+\eta}\right)^2} = \frac{1}{\eta(2+\eta)} \Rightarrow -\eta^2\sum_{i=1}^{k}\left(\frac{1}{1+\eta}\right)^{2i} = \frac{-\eta}{2+\eta}.
$$

Thus we have $x_\eta^* = \lim_{k\to\infty} x_k = \frac{-\eta}{2+\eta}$. Note we can also derive this from the consistency equation $x_\eta^* = \mathsf{PG}_g \circ \mathsf{PG}_f(x_\eta^*)$.

(b) For which combination above is the algorithm *consistent*, i.e. the limiting point $x_\eta^*$ of the algorithm equal to the true minimizer $x^* = \arg\min_{x\in\mathbb{R}} f(x) + g(x)$? Can you explain why only certain combinations are consistent?

**Solution:** Only for (ii) $\mathsf{GD}_g \circ \mathsf{PG}_f$ and (iii) $\mathsf{PG}_g \circ \mathsf{GD}_f$ is the limiting point of the algorithm equal to the true minimizer: $x_\eta^* = x^* = 0$. This is because the composition of the two operators in either (ii) or (iii) preserves $x^*$:

$$x^* = \mathsf{GD}_g \circ \mathsf{PG}_f(x^*)$$

$$x^* = \mathsf{PG}_g \circ \mathsf{GD}_f(x^*).$$

Indeed these are true for any convex functions $f$ and $g$, by the following argument:
Since $x^*$ minimizes $f + g$, it satisfies $\nabla f(x^*) + \nabla g(x^*) = 0$, or $\nabla f(x^*) = -\nabla g(x^*)$.
Consider (iii) $\mathsf{PG}_g \circ \mathsf{GD}_f$. The first step $\mathsf{GD}_f$ takes $x^*$ to $\mathsf{GD}_f(x^*) = x^* - \eta\nabla f(x^*) = x^* + \eta\nabla g(x^*)$. The second step $\mathsf{PG}_g$ takes $\mathsf{GD}_f(x^*)$ to $y^* = \mathsf{PG}_g(\mathsf{GD}_f(x^*))$ which (by definition) satisfies the optimality condition

$$y^* = \mathsf{GD}_f(x^*) - \eta\nabla g(y^*).$$

Rearranging and plugging in the form of $\mathsf{GD}_f(x^*)$:

$$y^* + \eta\nabla g(y^*) = \mathsf{GD}_f(x^*) = x^* + \eta\nabla g(x^*).$$

This shows that $y^* = \mathsf{PG}_g(\mathsf{GD}_f(x^*))$ is a solution, and it must be unique since it is the proximal operator of a convex function $g$. This shows that $x^*$ is preserved by the update (iii):

$$x^* = \mathsf{PG}_g(\mathsf{GD}_f(x^*)).$$

You can similarly show that for (ii): $x^* = \mathsf{GD}_g(\mathsf{PG}_f(x^*))$. (And you can check that this is not true for (i) and (iv).)
This means that $x^*$ is a fixed point to the updates in (ii) and (iii). We also know the limit $x_\eta^*$ of the algorithm is also a fixed point to the update. The limit point of the algorithm must be unique (because each step is a contraction map, since $f$ and $g$ are convex). Hence, the limit point $x_\eta^*$ is the same as the true minimizer $x^*$.

(P3) Let $X \sim \mathcal{N}(0, C)$ be a Gaussian random variable on $\mathbb{R}^d$ with mean $0 \in \mathbb{R}^d$ and covariance matrix $C \in \mathbb{R}^{d\times d}$. Assume $C \succ 0$ has eigenvalues $0 < \lambda_1 \le \cdots \le \lambda_d$. Evaluate the integral to compute the values below.

(a) Write down the following expression as a function of $\lambda_1, \ldots, \lambda_d$:

$$\int_{\mathbb{R}^d} e^{-\frac{1}{2}x^\top C^{-1}x} dx.$$

**Solution:** Recall density of Gaussian distribution $\mathcal{N}(0, C)$ is $\gamma(x) = \frac{1}{\sqrt{\det(2\pi C)}} e^{-\frac{1}{2}x^\top C^{-1}x}$. This must integrate to 1, so

$$\int_{\mathbb{R}^d} e^{-\frac{1}{2}x^\top C^{-1}x} dx = \sqrt{\det(2\pi C)} = (2\pi)^{d/2}\sqrt{\det C} = (2\pi)^{d/2}\sqrt{\prod_{i=1}^{d} \lambda_i}.$$

In the above, we use the fact that $\det(\alpha C) = \alpha^d \det C$ for a $d \times d$ matrix $C$ and $\alpha \in \mathbb{R}$. We also use the fact that the determinant of a matrix is the product of its eigenvalues: $\det(X) = \prod_{i=1}^{d} \lambda_i$.

(b) For $\theta \in \mathbb{R}^d$, compute $\mathbb{E}[e^{\theta^\top X}]$. When is it finite?

**Solution:** By completing the square inside the exponential, we can compute:

$$\begin{aligned}
\mathbb{E}[e^{\theta^\top X}] &= \frac{1}{\sqrt{\det(2\pi C)}} \int_{\mathbb{R}^d} e^{\theta^\top x} e^{-\frac{1}{2}x^\top C^{-1}x} dx \\
&= \frac{1}{\sqrt{\det(2\pi C)}} \int_{\mathbb{R}^d} \exp\left\{-\frac{1}{2}\left(x^\top C^{-1}x - 2\theta^\top x + \theta^\top C\theta\right) + \frac{1}{2}\theta^\top C\theta\right\} dx \\
&= \frac{1}{\sqrt{\det(2\pi C)}} \int_{\mathbb{R}^d} \exp\left\{-\frac{1}{2}\left((x - C\theta)^\top C^{-1}(x - C\theta)\right) + \frac{1}{2}\theta^\top C\theta\right\} dx \\
&= e^{\frac{1}{2}\theta^\top C\theta} \cdot \underbrace{\frac{1}{\sqrt{\det(2\pi C)}} \int_{\mathbb{R}^d} \exp\left\{-\frac{1}{2}\left((x - C\theta)^\top C^{-1}(x - C\theta)\right)\right\} dx}_{\int \mathcal{N}(x; C\theta, C)dx = 1} \\
&= e^{\frac{1}{2}\theta^\top C\theta}
\end{aligned}$$

This is finite for any $\theta \in \mathbb{R}^d$. (The function above is the moment generating function for a multivariate Gaussian.)

(c) For $t > 0$, compute $\mathbb{E}[e^{t\|X\|^2}]$. When is it finite?

**Solution:**

$$\begin{aligned}
\mathbb{E}[e^{t\|X\|^2}] &= \frac{1}{\sqrt{\det(2\pi C)}} \int_{\mathbb{R}^d} e^{t \cdot x^\top x} e^{-\frac{1}{2}x^\top C^{-1}x} dx \\
&= \frac{1}{\sqrt{\det(2\pi C)}} \int_{\mathbb{R}^d} e^{-\frac{1}{2}x^\top (C^{-1} - 2tI)x} dx
\end{aligned}$$

$$= \frac{\sqrt{\det\left(2\pi(C^{-1} - 2tI)^{-1}\right)}}{\sqrt{\det(2\pi C)}}.$$

In the last step above, we have used part (a), which is valid as long as $(C^{-1} - 2tI) \succ 0$. Since the eigenvalues of $C$ are $\lambda_i$, the eigenvalues of $C^{-1} - 2tI$ are $1/\lambda_i - 2t$. We want this to be positive: $1/\lambda_i - 2t > 0 \Rightarrow t < 1/(2\lambda_i)$ for all $i$. So $t < 1/(2\lambda_d)$ where recall $\lambda_d$ is the maximum eigenvalue of $C$.

Assume $t < 1/(2\lambda_d)$. We can further simplify the result above as:

$$\begin{aligned}
\mathbb{E}[e^{t\|X\|^2}] &= \frac{\sqrt{\det\left(2\pi(C^{-1} - 2tI)^{-1}\right)}}{\sqrt{\det(2\pi C)}} \\
&= \sqrt{\det(C^{-1}(C^{-1} - 2tI)^{-1})} \\
&= \sqrt{\det(I - 2tC)^{-1}} \\
&= \frac{1}{\sqrt{\prod_{i=1}^{d}(1 - 2t\lambda_i)}}.
\end{aligned}$$

<u>Note:</u> For a symmetric matrix $M \in \mathbb{R}^{d\times d}$, the integral $\int_{\mathbb{R}^d} e^{-\frac{1}{2}x^\top M x} dx$ is finite if and only if all its eigenvalues are positive, or $M \succ 0$. You can already see this in $d = 1$: $\int_{\mathbb{R}} e^{-\frac{m}{2}x^2} dx < \infty$ if and only if $m > 0$. In dimension $d$, we can use eigendecomposition to reduce the $d$-dimensional integral into a product of 1-dimensional integrals, so the conclusion holds if and only if all the eigenvalues of $M$ are positive, or $M \succ 0$.

(d) Compute the (negative) entropy $H(\rho) = \mathbb{E}[\log \rho(X)]$. How does it scale with $C$?

**Solution:**

$$\begin{aligned}
\mathbb{E}[\log \rho(X)] &= \mathbb{E}\left[\underbrace{-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det C}_{\text{const.}} - \frac{1}{2}x^\top C^{-1} x\right] \\
&= -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det C - \frac{1}{2}\mathbb{E}\left[X^\top C^{-1} X\right]
\end{aligned}$$

To simplify the last term:

$$\begin{aligned}
\mathbb{E}\left[X^\top C^{-1} X\right] &= \mathbb{E}\left[\mathsf{Tr}\left(X^\top C^{-1} X\right)\right] && \text{Trace trick for quadratic form} \\
&= \mathbb{E}\left[\mathsf{Tr}\left(C^{-1} X X^\top\right)\right] && \text{Cyclic property of trace} \\
&= \mathsf{Tr}\left(\mathbb{E}\left[C^{-1} X X^\top\right]\right) && \text{Linearity of } \mathbb{E}, \mathsf{Tr} \\
&= \mathsf{Tr}\left(C^{-1}\mathbb{E}[X X^\top]\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathsf{Tr}\left(C^{-1}C\right) \qquad\qquad (\mathbb{E}[X]=0 \Rightarrow \mathbb{E}[XX^\top] = \mathsf{Cov}(X) = C) \\
&= \mathsf{Tr}(I_d) \\
&= d
\end{aligned}
$$

Altogether,

$$H(\rho) = -\frac{d}{2}(\log(2\pi)+1) - \frac{1}{2}\log\det C.$$

We see that $H(\rho)$ is inversely proportional to $C$: If we increase $C$ (e.g. multiply $C$ by 2), then $H(\rho)$ will decrease. (Recall in this problem $H(\rho) = \mathbb{E}[\log\rho]$, which is the negative of the usual definition of entropy $-\mathbb{E}[\log\rho]$; the latter is increasing if we increase $C$.)

(P4) Consider the noisy recursion:

$$x_{k+1} = (1-\eta)x_k + \epsilon z_k$$

where $\eta > 0$ is step size and $\epsilon > 0$ is noise scale (usually $\eta, \epsilon \ll 1$), and $z_k \sim \mathcal{N}(0, I)$ an independent Gaussian random variable in $\mathbb{R}^d$. We start from any $x_0 \sim \rho_0$ to get $x_k \sim \rho_k$.

(a) Compute the mean $m_k = \mathbb{E}_{\rho_k}[x_k]$ and covariance matrix $C_k = \mathsf{Cov}_{\rho_k}(x_k)$ as a function of $m_0, C_0, k$. Determine how fast they converge to the limit $m_\infty = \lim_{k\to\infty} m_k$ and $C_\infty = \lim_{k\to\infty} C_k$.

**Solution:**

$$
\begin{aligned}
x_1 &= (1-\eta)x_0 + \epsilon z_0 \\
x_2 &= (1-\eta)x_1 + \epsilon z_1 = (1-\eta)^2 x_0 + (1-\eta)\epsilon z_0 + \epsilon z_1 \\
x_3 &= (1-\eta)x_2 + \epsilon z_3 = (1-\eta)^3 x_0 + (1-\eta)^2\epsilon z_0 + (1-\eta)\epsilon z_1 + \epsilon z_3 \\
&\vdots \\
x_k &= (1-\eta)^k x_0 + \sum_{i=0}^{k-1} \epsilon(1-\eta)^i \cdot z_{k-1-i}
\end{aligned}
$$

Recall that for two standard normal random variables, say $Z, Z'$, the combination $aZ+bZ'$ results in a random variable distributed normal with mean 0 and variance $a^2+b^2$ (and the multivariate version is $a^2I + b^2I$). So the term

$$\sum_{i=0}^{k-1} \epsilon(1-\eta)^i \cdot z_i \sim \mathcal{N}\left(0, \epsilon^2 \sum_{i=0}^{k-1}(1-\eta)^{2i} \cdot I_d\right),$$

where the variance formula can be simplified using partial sum formula $\epsilon^2 \sum_{i=0}^{k-1}(1-\eta)^{2i} = \epsilon^2 \cdot \frac{1-(1-\eta)^{2k}}{\eta(2-\eta)}$.

Then

$$x_k = (1-\eta)^k x_0 + \tilde{z}, \qquad \tilde{z} \sim \mathcal{N}\left(0, \epsilon^2 \cdot \frac{1-(1-\eta)^{2k}}{\eta(2-\eta)} \cdot I\right).$$

Note that $m_k = (1-\eta)^k m_0$ and $C_k = (1-\eta)^{2k} C_0 + \epsilon^2 \left(\frac{1-(1-\eta)^{2k}}{\eta(2-\eta)}\right) \cdot I$.

(b) Determine what is the limiting distribution $\pi^* = \lim_{k\to\infty} \rho_k$ of the recursion. How does it depend on the step size and noise scale?

**Solution:** In the above, note that as $k \to \infty$, the infinite series converges to $\frac{1}{\eta(2-\eta)}$ for $(1-\eta)^2 < 1 \iff -1 < \eta < 1$ (but $\eta > 0$ so $0 < \eta < 1$) and the $x_0$ term goes to zero so

$$x_\infty \sim \mathcal{N}\left(0, \frac{\epsilon^2}{\eta(2-\eta)} \cdot I_d\right) = \pi^*$$

which is a centered multivariate Gaussian.

(c) Suppose $\epsilon = \eta$. What happens to the limiting distribution $\pi^*$ for small $\eta \to 0$?
**Solution:** The limiting variance is

$$\lim_{\eta\to 0} \frac{\eta^2}{\eta(2-\eta)} = \frac{\eta}{2-\eta} = 0.$$

We approach point mass (Dirac delta) at 0.

(d) Suppose $\epsilon = \sqrt{\eta}$. What happens to the limiting distribution $\pi^*$ for small $\eta \to 0$?

**Solution:** The limiting covariance is

$$\lim_{\eta\to 0} \frac{\eta}{\eta(2-\eta)} = \frac{1}{2-\eta} = \frac{1}{2}.$$

The covariance matrix approaches $\frac{1}{2}I$ and the limiting distribution approaches $\mathcal{N}(0, \frac{1}{2}I)$.

(P5) (a) Recall the one-dimensional integration by parts formula (you may consult textbooks).

**Solution:** For differentiable functions $u, v\colon \mathbb{R} \to \mathbb{R}$ with $\lim_{x\to\pm\infty} u(x) = \lim_{x\to\pm\infty} v(x) = 0$,

$$\int_{-\infty}^{\infty} u(x)\, v'(x)\, dx = -\int_{-\infty}^{\infty} u'(x)\, v(x)\, dx.$$

(b) Use the formula above to prove the following multi-dimensional integration by parts identity. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a differentiable function with $\lim_{\|x\|\to\infty} f(x) = 0$. Let $v\colon \mathbb{R}^d \to \mathbb{R}^d$ be a differentiable vector field with $\lim_{\|x\|\to\infty} \|v(x)\| = 0$. Show that:

$$\int_{\mathbb{R}^d} \langle \nabla f(x), v(x) \rangle \, dx = -\int_{\mathbb{R}^d} f(x) \, \nabla \cdot v(x) \, dx \tag{1}$$

**Solution:** We will show that for each $i = 1, \ldots, d$:

$$\int_{\mathbb{R}^d} \frac{\partial f(x)}{\partial x_i} \, v_i(x) \, dx \overset{(*)}{=} -\int_{\mathbb{R}^d} f(x) \, \frac{\partial v_i(x)}{\partial x_i} \, dx.$$

We use the one-dimensional integration by parts from part (a). For each $i = 1, \ldots, d$, and for each $x_{\setminus i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) \in \mathbb{R}^{d-1}$:

$$\int_{-\infty}^{\infty} \frac{\partial f(x)}{\partial x_i} \, v_i(x) \, dx_i = -\int_{-\infty}^{\infty} f(x) \, \frac{\partial v_i(x)}{\partial x_i} \, dx_i.$$

Then by integrating over $x_{\setminus i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) \in \mathbb{R}^{d-1}$ and using Fubini's theorem to exchange the order of integration (assuming the functions are absolutely integrable):

$$\begin{aligned}
\int_{\mathbb{R}^d} \frac{\partial f(x)}{\partial x_i} \, v_i(x) \, dx &= \int_{\mathbb{R}^{d-1}} \left( \int_{-\infty}^{\infty} \frac{\partial f(x)}{\partial x_i} \, v_i(x) \, dx_i \right) dx_{\setminus i} \\
&= \int_{\mathbb{R}^{d-1}} \left( -\int_{-\infty}^{\infty} f(x) \, \frac{\partial v_i(x)}{\partial x_i} \, dx_i \right) dx_{\setminus i} \\
&= -\int_{\mathbb{R}^d} f(x) \, \frac{\partial v_i(x)}{\partial x_i} \, dx.
\end{aligned}$$

Then the desired result (1) follows by summing over $i = 1, \ldots, d$:

$$\begin{aligned}
\int_{\mathbb{R}^d} \langle \nabla f(x), v(x) \rangle \, dx &= \int_{\mathbb{R}^d} \sum_{i=1}^{d} \frac{\partial f(x)}{\partial x_i} \, v_i(x) \, dx \\
&= \sum_{i=1}^{d} \int_{\mathbb{R}^d} \frac{\partial f(x)}{\partial x_i} \, v_i(x) \, dx \\
&= -\sum_{i=1}^{d} \int_{\mathbb{R}^d} f(x) \, \frac{\partial v_i(x)}{\partial x_i} \, dx \\
&= -\int_{\mathbb{R}^d} f(x) \sum_{i=1}^{d} \frac{\partial v_i(x)}{\partial x_i} \, dx \\
&= -\int_{\mathbb{R}^d} f(x) \, \nabla \cdot v(x) \, dx.
\end{aligned}$$

□

(c) Let $X \sim \mathcal{N}(0, I)$ be a Gaussian random variable in $\mathbb{R}^d$. Prove **Stein's identity**:

$$\mathbb{E}[\nabla f(X)] = \mathbb{E}[X\, f(X)]$$

If $X \sim \mathcal{N}(m, C)$ for some $m \in \mathbb{R}^d$, $C \succ 0$, how does the identity above change?

**Solution:** Recall the density of the $\mathcal{N}(\mu, C)$ distribution is

$$\rho(x) = (\det(2\pi C))^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\top C^{-1}(x-\mu)\right).$$

Its gradient is

$$\nabla\rho(x) = -C^{-1}(x-\mu)\,\rho(x).$$

Then by the integration by parts identity,

$$\begin{aligned} C^{-1}\mathbb{E}[(X-\mu)\,f(X)] &= \int_{\mathbb{R}^d} \rho(x)\,C^{-1}(x-\mu)\,f(x)\,dx \\ &= -\int_{\mathbb{R}^d} \nabla\rho(x)\,f(x)\,dx \\ &\overset{(*)}{=} \int_{\mathbb{R}^d} \rho(x)\,\nabla f(x)\,dx \\ &= \mathbb{E}[\nabla f(X)]. \end{aligned}$$

Note that in step (*) above, we apply the one-dimensional integration by parts identity for each component $\frac{\partial\rho(x)}{\partial x_i}$ and $\frac{\partial f(x)}{\partial x_i}$ of $\nabla\rho(x)$ and $\nabla f(x)$. □

## Additional questions for 586

(Q1) Let $0 < \epsilon \ll 1$ and $y$ be a second-order perturbation of $x \in \mathbb{R}^d$ for some $u, v \in \mathbb{R}^d$:

$$y = x + \epsilon u + \epsilon^2 v$$

(a) Compute $\|y\|^2 - \|x\|^2$ as a polynomial in $\epsilon$.

**Solution:** We can compute:

$$\|y\|^2 = \|x\|^2 + \epsilon^2\|u\|^2 + \epsilon^4\|v\|^2 + 2\epsilon x^\top u + 2\epsilon^2 x^\top v + 2\epsilon^3 u^\top v.$$

Therefore,

$$\|y\|^2 - \|x\|^2 = 2\epsilon x^\top u + \epsilon^2(\|u\|^2 + 2x^\top v) + 2\epsilon^3 u^\top v + \epsilon^4\|v\|^2.$$

(b) Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be three-times differentiable. Compute $f(y) - f(x)$ up to $O(\epsilon^3)$ terms: $f(y) = f(x) + a(x)\epsilon + b(x)\epsilon^2 + O(\epsilon^3)$. Compute $a(x)$, $b(x)$ in terms of derivatives of $f(x)$.

**Solution:** We use Taylor expansion (at $y$ about $x$):

$$\begin{aligned}
f(y) &= f(x + \epsilon u + \epsilon^2 v) \\
&\approx f(x) + \langle \nabla f(x), (y-x) \rangle + \left\langle \frac{1}{2}\nabla^2 f(x), (y-x)(y-x)^\top \right\rangle + O(\epsilon^3) \\
&= f(x) + \langle \nabla f(x), \epsilon u + \epsilon^2 v \rangle + \left\langle \frac{1}{2}\nabla^2 f(x), \epsilon^2 (u+\epsilon v)(u+\epsilon v)^\top \right\rangle + O(\epsilon^3)
\end{aligned}$$

The second term simplifies:

$$\langle \nabla f(x), \epsilon u + \epsilon^2 v \rangle = \epsilon \nabla f(x)^\top u + \epsilon^2 \nabla f(x)^\top v$$

Simplifying the last term separately:

$$\begin{aligned}
&\left\langle \frac{1}{2}\nabla^2 f(x), \epsilon^2 (u+\epsilon v)(u+\epsilon v)^\top \right\rangle \\
&= \sum_{j=1}^{d}\sum_{i=1}^{d} \frac{1}{2}\frac{\partial^2}{\partial x_i \partial x_j} f(x) \cdot \epsilon^2 (u_i + \epsilon v_i)(u_j + \epsilon v_j) \\
&= \frac{\epsilon^2}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} f(x) \left( u_i u_j + \epsilon u_i v_j + \epsilon u_j v_i + \epsilon^2 v_i v_j \right) \\
&= \frac{\epsilon^2}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} f(x) \cdot u_i u_j + \frac{\epsilon^3}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} f(x) \cdot (u_i v_j + u_j v_i) + \frac{\epsilon^4}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} f(x) \cdot (v_i v_j) \\
&= \frac{\epsilon^2}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} f(x) \cdot u_i u_j + O(\epsilon^3) \\
&= \frac{\epsilon^2}{2}\left\langle \nabla^2 f(x), u u^\top \right\rangle + O(\epsilon^3)
\end{aligned}$$

Thus, we can write $f(y) - f(x)$ as

$$f(y) - f(x) = \Bigg( \underbrace{\nabla f(x)^\top u}_{a(x)} \Bigg) \cdot \epsilon + \Bigg( \underbrace{\nabla f(x)^\top v + \frac{1}{2}\left\langle \nabla^2 f(x), u u^\top \right\rangle}_{b(x)} \Bigg) \cdot \epsilon^2 + O(\epsilon^3).$$

(c) Let $\tilde{y} = x + \epsilon u$. Compute $f(\tilde{y}) - f(x)$ as a function of $\epsilon$, and compare with $f(y) - f(x)$.

**Solution:**

$$f(\tilde{y}) \approx f(x) + \langle \nabla f(x), \tilde{y} - x \rangle + \left\langle \frac{1}{2}\nabla^2 f(x), (\tilde{y}-x)(\tilde{y}-x)^\top \right\rangle + O(\epsilon^3)$$

$$
\begin{aligned}
&= f(x) + \langle \nabla f(x), \epsilon u \rangle + \left\langle \frac{1}{2}\nabla^2 f(x), \epsilon^2 uu^\top \right\rangle + O(\epsilon^3) \\
&= f(x) + (\nabla f(x)^\top u) \cdot \epsilon + \left( \frac{1}{2} \left\langle \nabla^2 f(x), uu^\top \right\rangle \right) \cdot \epsilon^2 + O(\epsilon^3) \\
&\Rightarrow f(\tilde{y}) - f(x) = (\nabla f(x)^\top u) \cdot \epsilon + \left( \frac{1}{2} \left\langle \nabla^2 f(x), uu^\top \right\rangle \right) \cdot \epsilon^2 + O(\epsilon^3)
\end{aligned}
$$

The first term is the same with $f(y) - f(x)$, but the last term differs by $\nabla f(x)^\top v \cdot \epsilon^2$. This makes sense, since the perturbation $v$ happens at the $\epsilon^2$ scale.

(d) Suppose we want to minimize $f$ and we choose $u = -\nabla f(x)$, so up to first-order we are following gradient descent. What $v$ should we choose to further decrease $f(y)$?

**Solution:** Choose $v = -\nabla f(x)$ (or any positive multiple of it). The inner product $\nabla f(x)^\top v$ is negative for any vector $v$ along direction $-\nabla f(x)$.

(Q2) Consider the recursion

$$x_{k+1} = x_k - \eta \nabla f(x_k) + b$$

where $0 < \eta \leq \frac{1}{L}$ is step size and $b \in \mathbb{R}^d$. Assume $f$ is $\alpha$-strongly convex and $L$-smooth for some $\alpha > 0$, $L < \infty$.

(a) Show the map $F_\eta(x) = x - \eta \nabla f(x) + b$ is contractive: $\|F_\eta(x) - F_\eta(y)\| \leq (1 - \alpha\eta)\|x - y\|$.

**Solution:** This is because the map $F_\eta$ is $(1 - \alpha\eta)$-Lipschitz, which we can see because the Jacobian matrix of $F_\eta$ has eigenvalues bounded by $1 - \eta\alpha$.
For any $x, y \in \mathbb{R}^d$:

$$
\begin{aligned}
F_\eta(x) - F_\eta(y) &= (x - \eta\nabla f(x) + b) - (y - \eta \nabla f(y) + b) \\
&= (x - y) - \eta(\nabla f(x) - \nabla f(y))
\end{aligned}
$$

Note since $f$ is $\alpha$-strongly convex and $L$-smooth, $\alpha I \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^d$. Since

$$\nabla F_\eta(x) = I - \eta \nabla^2 f(x)$$

We have

$$0 \preceq (1 - \eta L)I \preceq \nabla F_\eta(x) \preceq (1 - \eta\alpha)I.$$

This means the map $F_\eta$ is $(1 - \alpha\eta)$-Lipschitz, since the Jacobian matrix $\nabla F_\eta$ has all eigenvalues bounded by $1 - \eta\alpha$: $\|\nabla F_\eta(x)\|_{\mathsf{op}} \leq 1 - \eta\alpha$.

Concretely, we can argue as follows. Let us interpolate from $x(0) = x$ to $x(1) = y$ via linear map $x(t) = (1-t)x + ty$ for $0 \le t \le 1$, so $\dot{x}(t) = y - x$. Then we can write

$$\begin{aligned} F_\eta(y) - F_\eta(x) &= F_\eta(x(1)) - F_\eta(x(0)) \\ &= \int_0^1 \frac{d}{dt} F_\eta(x(t))\, dt \\ &= \int_0^1 \nabla F_\eta(x(t)) \dot{x}(t)\, dt \\ &= \int_0^1 \nabla F_\eta(x(t))(y-x)\, dt. \end{aligned}$$

Therefore,

$$\begin{aligned} \|F_\eta(y) - F_\eta(x)\| &= \left\| \int_0^1 \nabla F_\eta(x(t))(x-y) dt \right\| \\ &\le \int_0^1 \|\nabla F_\eta(x(t))(x-y)\| dt \\ &\le \int_0^1 \|\nabla F_\eta(x(t))\|_{\mathsf{op}} \cdot \|x-y\| dt \\ &\le (1-\alpha\eta)\|x-y\|. \end{aligned}$$

(b) Show that there is a unique limit point $x_\infty \in \mathbb{R}^d$ of $F_\eta$ and that $x_k \to x_\infty$ exponentially fast: $\|x_k - x_\infty\| \le e^{-\alpha\eta k}\|x_0 - x_\infty\|$ for all $k \ge 0$. Compute $x_\infty$ in terms of $b$, $f$, $\eta$.

**Solution:** Note since the map $F_\eta$ is a contraction, it must have a limit point $x_\infty = \lim_{k\to\infty} x_k$, which is a fixed point of the map: $F_\eta(x_\infty) = x_\infty$. Note the limit must be unique, because if we have two fixed points $x_\infty, y_\infty$ of $F_\eta$, then after one step of $F_\eta$,

$$\|x_\infty - y_\infty\| = \|F_\eta(x_\infty) - F_\eta(y_\infty)\| \le (1-\eta\alpha)\|x_\infty - y_\infty\|.$$

This implies $\|x_\infty - y_\infty\| = 0$, so $x_\infty = y_\infty$.

From any $x_0$, $x_k$ converges to $x_\infty$ exponentially fast. This is because in each step, the distance decreases by a constant factor less than 1:

$$\|x_{k+1} - x_\infty\| = \|F_\eta(x_k) - F_\eta(x_\infty)\| \le (1-\eta\alpha)\|x_k - x_\infty\|.$$

Therefore, after $k$ iterations,

$$\|x_k - x_\infty\| \le (1-\eta\alpha)^k \|x_0 - x_\infty\|.$$

Since $1 - \alpha\eta \le e^{-\alpha\eta}$, we can further bound this as

$$\|x_k - x_\infty\| \le e^{-\alpha\eta k}\|x_0 - x_\infty\|.$$

To solve for $x_\infty$, note that

$$\begin{aligned} x_\infty &= F_\eta(x_\infty) \\ \Leftrightarrow \quad x_\infty &= x_\infty - \eta\nabla f(x_\infty) + b \\ \Rightarrow \nabla f(x_\infty) &= \frac{1}{\eta}b. \end{aligned}$$

Thus, the limit $x_\infty$ is the solution to $\nabla f(x_\infty) = \frac{1}{\eta}b$.
Abstractly, we want to invert the gradient operator: $x_\infty = (\nabla f)^{-1}(\frac{1}{\eta}b)$. Recall from convex analysis this is achieved by the gradient of the dual function: $(\nabla f)^{-1} = \nabla f^*$ (which means $y = \nabla f(x)$ if and only if $x = \nabla f^*(y)$). Here $f^*(y) = \sup_x \langle x, y\rangle - f(x)$ is the dual function (convex conjugate) of $f$, and recall the gradient is the maximizer: $\nabla f^*(y) = \arg\max_{x\in\mathbb{R}^d} \langle x, y\rangle - f(x)$.
This means we can write the limit $x_\infty$ above as:

$$x_\infty = \nabla f^*\left(\frac{1}{\eta}b\right) = \arg\max_{x\in\mathbb{R}^d} \frac{1}{\eta}\langle x, b\rangle - f(x).$$

(c) Let $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$. Give an upper bound to $\|x_\infty - x^*\|$ in terms of $b, \alpha, \eta$.

**Solution:** Since $\nabla f(x^*) = 0$, we can compute:

$$\|\nabla f(x_\infty) - \nabla f(x^*)\| = \|\nabla f(x_\infty)\| = \|\frac{1}{\eta} \cdot b\| = \frac{\|b\|}{\eta}.$$

Since $f$ is $\alpha$-strongly convex, we also have

$$\alpha\|x_\infty - x^*\| \le \|\nabla f(x_\infty) - f(x^*)\| \le \frac{\|b\|}{\eta}$$

This implies the bound

$$\|x_\infty - x^*\| \le \frac{\|b\|}{\alpha\eta}.$$

(Q3) Describe your research. What is the problem? How does it relate to probabilistic modeling or inference? (If you don't have research experience, you may describe a topic from a paper or a book.)